e-Informatica Software Engineering Journal, Volume 5, Issue 1, 2011, pages: 25–37, DOI 10.2478/v10233-011-0028-y

VERSITA

Examining Correlations in Usability Data to Effectivize Usability Testing

Jeff Winter*, Mark Hinely**
*School of Computing, Blekinge Institute of Technology
**, UIQ Technology AB
jeff.winter@bth.se, mark.hinely@bredband.net

Abstract

Based on a case study performed in industry, this work deals with a statistical analysis of data collected during usability testing. The data is from tests performed by usability testers from two companies in two different countries. One problem in the industrial situation is the scarcity of testing resources, and a need to use these resources in the most efficient way. Therefore, the data from the testing is analysed to see whether it is possible to measure usability on the basis of one single metric, and whether it is possible to judge usability problems on the basis of the distribution of use case completion times. This would allow test leaders to concentrate on situations where there are obvious problems. We find that it is not possible to measure usability through the use of one metric, but that it may be possible to gain indications of usability problems on the basis of an analysis of time taken to perform use cases. This knowledge would allow the collection of usability data from distributed user groups, and a more efficient use of scarce testing resources.

1. Introduction

The background to this study is the situation faced by companies developing and testing consumer products for a mass market. The study is based on a long research cooperation between Blekinge Institute of Technology (BTH) and UIQ Technology AB (UIQ), an international company established in 1999. UIQ, who developed and licensed a user interface platform for mobile phones, identified a need to develop a flexible test method for measuring the usability of mobile phones, to give input to design and development processes, and to present usability findings for a number of stakeholders at different levels in the organization. This need resulted in the development of UIQ Technology Usability Metrics (UTUM). UTUM was successfully used in operations at UIQ until the closure of the company in 2009.

Together with UIQ we found that there is a need for methods that can simplify the discovery of usability problems in mobile phones. There is also a desire to find ways of identifying usability problems in phones without having to engage the test leader in every step of the process, with the ability to do it for geographically dispersed user groups. However, we also realise that even if it is found to be possible to identify problem areas, for example through a simple measurement of one metric, or through an analysis of completion times, this would not identify the particular aspects of the use cases that are problematic for the users. It would simply indicate use cases where the users experienced problems. This means that further studies would still have to be performed by test leaders together with users, to examine and understand what the actual problems consist of, and how they affect the way that users experience the use of the phone. This must be done in order to create design solutions to alleviate the problems.

As we discuss in greater detail in section 3 of this article, the role of the usability tester is central in many ways, and it is a role that is not easily filled. It demands particular personal qualities, knowledge and experience. It involves the ability to communicate with people on many organisational levels, the ability to observe, record and analyse the testing process, and the ability to present the results of testing to many different stakeholders. Since there is a scarcity of people who can fill this role, it would ease the situation for companies wanting to perform usability testing if these resources could be used in the most efficient way possible. This is the principle behind the need to identify problematic use cases without having to involve the test leader in every step of the process.

If it is possible to identify use cases that are problematic, without requiring the presence of the test leader, this will allow companies to pinpoint which areas require further testing, so that test leaders can work more efficiently. Since we are working with a mass-market product, being able to do this remotely, for widely dispersed groups, would also be an advantage for the company, in order to test solutions in different geographical areas without requiring the usability tester to travel to these areas before there is seen to be a need, and to reduce the amount of testing that needs to be done on-site.

These needs are the basis of this article. In this work, we examine the metrics collected in the UTUM testing, to study the correlations between the metrics for efficiency, effectiveness, and satisfaction, to see whether we can measure usability on the basis of one metric, and we examine whether it is possible to develop a simple method of automatically identifying problem areas simply by measuring and analysing the time taken to perform different use cases.

2. Research Questions

The aim of this study is to examine whether there is a simple measurement to express usability, and to find if it is possible to streamline the discovery of problematic use cases. To do this, we examine the correlation between metrics for efficiency, effectiveness and satisfaction that have been collected during the testing process. These are the different elements of usability as specified in ISO 9241-11:1998 [1],). This is done in order to see whether there are correlations that allow us to discover usability problems on the basis of a simple metric. To satisfy the needs within industry, this metric should preferably be one that can easily be measured without the presence of the test leader. Based on this situation, we have formulated two research questions:

- **RQ1**: What is the correlation between the different aspects of usability (Effectiveness, Efficiency and Satisfaction)?
- **RQ2**: Can a statistical analysis of task-completion time allow us to discover problematic use cases?

The first research question is based on the idea that there may be a sufficiently strong correlation between the 3 factors of usability that measuring one of them would give a reliable indication of the usability of a mobile phone. The second research question is based on the theory that there is an expected distribution of completion times for a given use case and that deviations from goodness of fit indicate user problems.

This study is a continuation of previous efforts to examine the correlations between metrics for efficiency, effectiveness and satisfaction. A previous study by Frøkjær et al [2] found only weak correlations between the different factors of usability, whereas a study by Sauro [3] showed stronger correlations between the different elements. The results of this study will be placed in relation to these studies, to extend knowledge in the field. This is also a continuation of our previous work, where we have examined how the UTUM test contributes to quality assurance, and how it balances the agile and plan-driven paradigms (see e.g. [4, 5, 6, 7]).

3. Usability and the UTUM Test

UTUM is an industrial application developed and evolved through a long term cooperation between BTH and UIQ. UTUM is a simple and flexible usability test framework grounded in usability theory and guidelines, and in industrial software engineering practice and experience.

According to ISO 9241-11:1998 [1], Usability is the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. Effectiveness is the accuracy and completeness with which users achieve specified goals. Efficiency concerns the resources expended in relation to the accuracy and completeness with which users achieve goals. Satisfaction concerns freedom from discomfort, and positive attitudes towards the use of the product.

UTUM measures the usability of products on a general level, as well as on a functional level. According to Hornbæk [8], amongst the challenges when measuring usability are to distinguish and compare subjective and objective measures of usability, to study correlations between usability measures as a means for validation, and to use both micro and macro tasks and corresponding measures of usability. Emphasis is also placed on the need to represent the entire construct of usability as a single metric, in order to increase the meaningfulness and strategic importance of usability data [3]. UTUM is an attempt to address some of these challenges.

An important characteristic of the UTUM test is the approach to understanding users and getting user input. Instead of simply observing use, a test expert interacts and works together with the users to gain insight into how they experience being a mobile phone user, in order to gain an understanding of the users' perspective. Therefore, users who help with UTUM testing are referred to as testers, because they are doing the testing, rather than being tested. The representative of the development company is referred to as the test leader, or test expert, emphasising the qualified role that this person assumes.

The test experts are specialists who bring in and communicate the knowledge that users have, in accordance with Pettichord [9], who claims that good testers think empirically in terms of observed behaviour, and must be encouraged to understand customers' needs. Evidence in Martin et al [10] suggests that drawing and learning from experience may be as important as taking a rational approach to testing. The fact that the test leaders involved in the testing are usability experts working in the field in their everyday work activities means that they have considerable experience of their products and their field. They have specialist knowledge, gained over a period of time through interaction with end-users, customers, developers, and other parties that have an interest in the testing process and results. However, these demands placed on the background and skills of test leaders mean that these types of resources are scarce, and must be used in the most efficient way possible.

A second characteristic of UTUM is making use of the inventiveness of phone users, by allowing users to participate actively in the design process. The participatory design tradition [11] respects the expertise and skills of the users, and this, combined with the inventiveness observed when users use their phones, means that users provide important input for system development. The test expert has an important role to play as an advocate and representative of the user perspective. Thus, the participation of the user provides designers, with the test expert as an intermediary, with good user input throughout the development process.

The user input gained through the testing is used directly in design and decision processes. Since the tempo of software development in the area of mobile phones is high, it is difficult to channel meaningful testing results to recipients at the right time in the design process. To address, this problem, the role of the test expert has been integrated into the daily design process. UTUM testing is usually performed in-house, and results of testing can be channelled to the most critical issues. The continual process of testing and informal relaying of testing results to designers leads to a short time span between discovering a problem and implementing a solution.

The results of testing are summarised in a clear and concise fashion that still retains a focus on understanding the user perspective, rather than simply observing and measuring user behaviour. The results of what is actually qualitative research are summarised by using quantitative methods. this gives decision makers results in the type of presentations they are used to dealing with. Statistical results are not based on methods that supplant the qualitative methods that are based on PD and ethnography, but are ways of capturing in numbers the users' attitudes towards the product they are testing.

A UTUM test does not take place in a laboratory environment, but should preferably take place in an environment that is familiar to the person who is participating in the test, in order that he or she should feel comfortable. When this is not possible, it should take place in an environment that is as neutral as possible. Although the test itself usually takes about 20 minutes, the test leader books one hour with the tester, in order to avoid creating an atmosphere of stress. The roles in testing are the test leader, who is usually a usability expert, and the tester.

In the test, the test leader welcomes the tester, and tries to put the tester at their ease. This includes explaining the purpose of the test, and saying that it is the telephone that is being tested, not the performance of the tester. The tester is instructed to tell the test leader when she or he is ready to begin the use case, so that the test leader can start the stopwatch to time the use case, and the tester should also tell the test leader when the use case is complete.

The tester begins by filling in some of their personal details and some general information about their phone usage. This includes name, age, gender, previous telephone use, and other data that can have an effect on the result of the test, such as which applications they find most important or useful. In some circumstances, this data can also be used to choose use cases for testing, based on the tester's use patterns.

For each phone to be tested, the tester is given time to get acquainted with the device. If several devices are to be tested, all of the use cases are performed on one device before moving on to the next phone. The tester is given a few minutes to get acquainted with the device, so that he or she can get a feeling for the look and feel of the phone. When this has been done, the tester fills in a Hardware Evaluation, a questionnaire based on the System Usability Scale (SUS) [12] about attitudes to the look and feel of the device. The SUS was developed in 1986 by John Brooke, then working at the Digital Equipment Company. The SUS consists of 10 statements, where even-numbered statements are worded negatively, and odd-numbered statements are worded positively.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The answers in the SUS are based on Likert style responses, ranging from "Strongly disagree" to "Strongly agree". The Likert scale is a widely used summated rating that is easy to develop and use. People often enjoy completing this type of scale, and they are likely to give considered answers and be more prepared to participate in this than in a test that they perceive as boring ([13] p. 293).

Brooke characterised the SUS as being a "Quick and Dirty" method of measuring usability. However, Lewis and Sauro state that although SUS may be quick, it is probably not dirty, and they cite studies that show that SUS has been found to be a reliable method of rating usability [14]. SUS has been widely used in the industrial setting, and Lewis and Sauro state that the SUS has stood the test of time, and they encourage practitioners using the SUS to continue to do so, and show how SUS can be decomposed into Usability and Learnability components, beyond showing the overall SUS score [14]. In a study of questionnaires for assessing the usability of a website, Tullis and Stetson found that the SUS, which was one of the simplest questionnaires studied, was found to yield amongst the most reliable results across sample sizes, and that SUS was the only questionnaire of those studied that addressed all of the aspects of the users' reactions to the website as a whole [15].

In a UTUM test, the users perform the use cases, the test leader observes what happens, and records the time taken to execute the tasks, observes hesitation or divergences from a natural flow use, notes errors, and counts the number of clicks to complete the task. Data is recorded in a form where the test leader can make notes of their observations. The test leader ranks the results of the use case on a scale between 0 - 4, where 4 is the best result. This judgement is based on the experience and knowledge of the test leader. This means that the result is not simply based on the time taken to perform the use case, but also on the flow of events, and events that may have affected the completion of the use case.

After performing each use case, the tester completes a Task Effectiveness Evaluation, a shortened SUS questionnaire [12] concerning the phone in relation to the specific use case performed. This is repeated for each use case. Between use cases, there is time to discuss what happened, and to explain why things happened the way they did. The test leader can discuss things that were noticed during the test, and see whether his or her impressions were correct, and make notes of comments and observations. Even though the test leader in our case does not usually actively probe the tester's understanding of what is being tested, this gives the opportunity to ask follow up questions if anything untoward occurs, and the chance to converse with the tester to glean information about what has occurred during the test.

The final step is an attitudinal metric representing the user's subjective impressions of how easy the phone is to use. This is found through the SUS [12], and it expresses the tester's opinion of the phone as a whole. The statements in the original SUS questionnaire are modified slightly, where the main difference is the replacement of the word "system" with the word "phone", to reflect the fact that a handheld device is being tested, rather than a system. This SUS questionnaire results in a number that expresses a measure of the overall usability of the phone as a whole. In general, SUS is used after the user has had a chance to use the system being evaluated, but before any debriefing or discussion of the test. In UTUM testing, the tester fills in the SUS form together with the test leader, giving an opportunity to discuss issues that arose during the test situation.

The data collected during the test situation is used to calculate a number of metrics, which are then used to make different presentations of the results to different stakeholders. These include the Task Effectiveness Metric, which is determined by looking at each use case and determining how well the telephone supports the user in carrying out each task. It is in the form of a response to the statement "This telephone provides an effective way to complete the given task". It is based on the test leader's judgement of how well the use case was performed, recorded in the test leader's record and the answers to the Task Effectiveness Evaluation. The Task Efficiency Metric is a response to the statement "This telephone is efficient for accomplishing the given task". This is calculated by looking at the distribution of times taken for each user to complete each use case. The distribution of completion times is used to calculate an average value for each device per use case. The User Satisfaction Metric, is calculated as an average score for the answers in the SUS, and is a composite response to the statement "This telephone is easy to use". For more information regarding different ways of presenting these metrics and data, see ([7], Appendix A).

A previous study by Winter et al [6] showed that two different groups of stakeholders existed within UIQ. The first group was designated as Designers represented by e.g. interaction designers and system and interaction architects, representing the shop floor perspective. The second group was designated as Product Owners, including management, product planning, and marketing, representing the management perspective. These two groups were found to have different needs regarding the presentation of test results. These differences concerned the level of detail included in the presentation, the ease with which the information can be interpreted, and the presence of contextual information included in the presentation. Designers prioritised presentations that gave specific information about the device and its features, whilst Product Owners prioritised presentations that gave more overarching information about the product as a whole, and that were not dependent on including contextual information.

These results, and more information on UTUM in general, are presented in greater detail in [7] (chapter 4 and Appendix A). A video demonstration of the test process (ca. 6 minutes) can be found on YouTube [16].

4. Research Method

The cooperative research and development work that led to the development of UTUM has been based on an action research approach according to the research and method development methodology called Cooperative Method Development (CMD) (see e.g. [17]). CMD is an approach to research that combines qualitative social science fieldwork, with problem-oriented method, technique and process improvement. CMD has as its starting point existing practice in industrial settings, and although it is motivated by an interest in use-oriented design and development of software, it is not specific for these methods, tools and processes.

This particular work is based on a case study [18] and grounded theory [19] approach. A case study is "an empirical enquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident" ([18], p. 13). The focus is on a particular case, taking the context into account, involving multiple methods of data collection; data can be both qualitative and quantitative, but qualitative data are almost always collected ([13] p. 178). Case studies have their basis in a desire to understand complex social phenomena, and are useful when "how" or "why" questions are being asked, and where the researcher has little control over events ([18], p. 7). A case study approach allows the retention of characteristics of real life events [18].

The data in this case study has been analysed in a grounded theory approach. Grounded theory (GT) is both a strategy for performing research and a style of analysing the data that arises from the research ([13], p. 191). It is a systematic but flexible research style that gives detailed descriptions for data analysis and generation of theory. It is applicable to a large variety of phenomena and is often interview-based and ([13], p. 90) but other methods such as observation and document analysis can also be used ([13], p. 191). We have not attempted to work according to pure GT practice, and have applied a case study perspective, using ethnography [20] and participatory design [11].

5. Subjects and Context

The data in this study were collected in tests performed by UIQ in Sweden and by a tester from a mobile phone manufacturer in England. The testing was performed in a situation where there are complex relationships between customers, clients, and end-users, and complexities of how and where results were to be used. The phones were a UIQ phone, a "Smart phone" of a competing brand, and a popular consumer phone. The use cases were decided by the English company, and were chosen from their 20 most important use cases for a certain mobile phone. The use cases were:

- UC1. Receive and answer an incoming call
- UC2. Save the incoming call as a new contact - "Joanne"
- UC3. Set an alarm for 8 o'clock tomorrow morning
- UC4. Read an incoming SMS and reply with "I'm fine"

- UC5. Make a phone call to Mårten (0708570XXX)
- UC6. Create a new SMS - "Hi meet at 5" and send to Joanne (0708570XXX)

The test group consisted of 48 testers. The group consisted of 24 testers from Sweden, and 24 testers from England, split into 3 age groups: 17 - 24; 25 - 34; 35+. Each age group consisted of 8 females and 8 males. The size of the group was in order to get results from a wide range of testers to obtain general views, and to enable comparisons between age groups, cultures and genders. Normally, it was not deemed necessary to include so many testers, as small samples have been found to be sufficient to evaluate products. Dumas and Reddish [21] for example, refer to previous studies that indicate in one case that almost half of all major usability problems were found with as few as three participants, and in a second case that a test with four to five participants detected 80% of usability problems, whilst ten participants detected 90% of all problems. This indicates that the inclusion of additional participants is less and less likely to contribute new information. The number of people to include in a test thus depends on how many user groups are needed to satisfy the test goals, the time and money allocated for the test, and the importance of being able to calculate statistical significance.

However, even though this can be seen from the point of view of the participating organisations as a large test, compared to their normal testing needs, where the data collected consisted of more than 10 000 data points, the testing was still found to be a process where results were produced quickly and efficiently. In this case, the intention of using a larger number of testers was to obtain a greater number of tests, to create a baseline for future validation of products, to identify and measure differences or similarities between two countries, and to identify issues with the most common use-cases. Testers were drawn from a database of mobile phone users who have expressed an interest in being testers, and who may or may not have been testers in previous projects.

6. Validity

Regarding internal reliability, the data used in this study have been collected according to a specified testing plan that has been developed over a long period of time, and that has been used and found to be a useful tool in many design and development projects. The risk of participant error in data collection is small, as the test is monitored, and the data is verified by the test leader. The risk of participant bias is also small, as the testers are normal phone users, using a variety of different phones, and they gain no particular benefits from participating in the tests or from rating one device as being better than another. The fact that much of the data has been in the form of self evaluations completed by the testers themselves, and that the testing has been performed by specialized usability experts minimizes the risk of observer error. The risk of observer bias is dealt with by the presence of the two independent test leaders, allowing us to compute inter-observer agreements. The use of multiple methods of data collection, including self assessment, test leader observation and measurement, and the collection of qualitative data, allow us to base our findings on many types and ways of collecting data.

In regard to external validity, the fact that the testing has been performed in two different countries may be seen as a risk, but the two countries, Sweden and England, are culturally relatively close, which should mean that the results are comparable across the national boundaries. The tasks performed in the testing are standard tasks that are common to most types of mobile phones, and should therefore not affect the performance or results of the tests. The users are a cross section of phone users, and the results should thus be generalisable to the general population of phone users.

To ensure the statistical conclusion validity, we use statistical methods that are standard in the field, and use the SPSS software package PASW Statistics 18 for statistical analysis.

7. Data Sets, Possible Correlations and Analysis Steps

The test data has been split into three sets of data. This division is based on the metrics collected in the attitudinal questionnaires and the times recorded by the test leader during testing. These data sets concern satisfaction, effectiveness and efficiency, as called for by ISO 9241-11:1998 [1]. The sets of test data are:

Set 1: SUSuapp - Based on the System Usability Scale (SUS) [12], which consists of 10, 5-scale Likert questions. The evaluation is a user appraisal of satisfaction, based on one evaluation per phone and tester. It is a summary of the use cases performed on the individual phones. It provides us with a total of 144 data points - 48 per phone (48 testers, 3 phones, 1 SUS per phone).

Set 2: TEEuapp - Based on a Task Effectiveness Evaluation (TEE), which consists of 6, 5-scale Likert questions. It is a user appraisal based on one evaluation per phone, use case and tester. The tester fills in this evaluation directly after completing each of the 6 use cases on each of the three phones. It provides us with a total of 864 data points - 144 per use case task (48 testers, 6 use cases, 3 phones).

Set 3: TIMEreal - This is used to represent efficiency, and is the time taken in seconds to complete a use case task. It is a test leader measurement based on one number per phone, use case and tester. The test leader measures the time for the tester to complete each of the use cases on each of the phones. This provides us with 864 data points- 144 per use case task (48 testers, 6 use cases, 3 phones).

As a complement to these data, we also make use of a spreadsheet, the Structured Data Summary (SDS) [22] that is used to record qualitative data based on the progress of the testing. This contains some of the qualitative findings of the testing and the SDS shows issues that have been found, for each tester, and each device, for every use case. Comments made by the testers and observations made by the test leader are stored as comments in the SDS.

The first step in the data analysis is to investigate the strength of the correlations between the metrics for satisfaction, effectiveness and efficiency. The second step is to investigate if the distribution of time taken to perform use cases can provide a reliable indication of problematic use cases, and in which way this should be analyzed and shown. If this is successful, it should be possible to discover use cases that exhibit poor usability by looking at the shape of the distribution curve. The third step is to verify the fact that the distribution of time can be used to illustrate the fact that certain use cases exhibit poor usability. This can be done by comparing with the data recorded in the SDS for these use cases, to see if the test leader has noted problems that users experienced. If this is found to be the case, this indication could be used when testing devices, to identify the areas where test leader resources should be directed, thus allowing a more efficient use of testing resources.

STEP 1: Investigate the correlation between Satisfaction, Effectiveness and Efficiency. For each phone each tester completed a SUS-evaluation (SUSuapp). SUSuapp gives an appraisal score from 0-40. The correlation between the SUSuapp, and TEEuapp might be calculated using Pearson's correlation coefficient, Spearman rank correlation coefficient, or Kendall's rank correlation coefficient (Kendall's Tau). The most reasonable method could be Spearman or Kendall's tau, as these deal with data in the form of ranks or ordering of data, and do not assume normal distribution of the data, on which the Pearson coefficient is based. Spearman is preferred by some, as it is in effect a Pearson coefficient performed on ranks, but Kendall's Tau is usually preferred, as it deals with ties more consistently [13]

The SUSuapp data is the result the 144 Likert appraisals, which could normally be assumed to exhibit a normal distribution. However, in some of the other data distributions, we have observed a positive skew that also suggests that Spearman may be a better choice. Also, the central concentration of the data causes many ties in ranks, which could make Kendall's Tau more appropriate.

The tests that include TIMEreal may be more difficult to deal with. Since the TIMEreal data is continuous, while the other data is of Likert-type, it may be difficult to see any linear relationships. However, the same tests should still be performed. The results of the analysis are found in Table 1.

The analysis shows only weak to moderate correlations between the different factors. This is particularly obvious regarding Kendall's tau, which as previously mentioned is probably the best indicator given the type of data involved here. This supports the findings of Frøkjær [2], who state that all three factors must be measured to gain a complete picture of usability. It contradicts the results of Sauro et al [3], who showed stronger correlations, although even Sauro et al state that it is important to measure all three factors, since each measure contains information not contained in the other measures.

These results do not support our conjecture that there is a sufficiently strong correlation between the 3 factors of usability that simply measuring one of them would give a reliable indication of the usability of a mobile phone.

STEP 2: Investigating if the distribution of time can provide a reliable indication of problematic use cases. We find that TIMEreal data, for time taken to complete a given use case, corresponds well with a Rayleigh distribution (Ray(2*mean)) with a shape parameter that is twice the mean of the data. Data points that end up in the tail fall under a specific degree of probability of belonging to the Ray(2*mean) distribution. This means that the use cases with a "long tail" are those that the testers found to be troublesome (see Fig 1).

Figure 1 illustrates one use case. The right hand diagram is the seconds to complete the use case divided into ten evenly spaced frequency intervals. The diagram to the left is the Ray(2*mean) probability distribution. For example, we see that 2 on the x-axis has a 28% chance belonging to the Rayleigh distribution, and that is where we have a frequency of 70+ data points. 6 on the x-axis has a less than 1% chance of belonging to the distribution and we see that 6 in our data is empty. This would mean that the points in our data set in ranges 7-10 are beyond all probability influenced by something more than the excepted random difference between different testers. Our interpretation is that these are the use cases where "something went wrong". This result suggests that it may be possible to discover use cases where users have problems, by examining the distribution of the time taken to perform the use case.

STEP 3: Verifying the "long tail" method of identifying troublesome use cases. Here we analyse which use cases the testers have experienced as exhibiting poor usability by analysing the distribution of time taken to complete the use case. This is cross tabulated with data from the SDS [22], the spreadsheet containing some of the qualitative findings of the testing. The SDS shows issues that have been found, for each tester, and each device, for every use case. Comments made by the testers and observations made by the test leader are stored as comments in the spreadsheet.

Given the fact that the intention of this work is to find ways that simplify the discovery of problematic use cases, and the fact that the test is designed to be flexible and simple to perform and analyse, we attempted to find some simple heuristic that could help us differentiate between the use cases with high and low levels of problems. We ordered the use cases according to their coefficients of variation, which is the standard deviation divided by the mean time taken to perform the use case. This allows us to standardize the use cases, in order to give a basis for comparison.

This calculation gave us a spread between 0.481 and 1.074. To give a simple cut-off point between Lower and Higher problem use cases, we set a boundary where a coefficient of variation of 0.6 is regarded as High problem, thus dividing the set of use cases into two groups. We also use a simple heuristic to judge an acceptable level of problems when performing a use case. The test leader registers problems observed whilst performing the use case by a letter "y" in the SDS, with an explanatory comment. One tester may have experienced more than one problem, and all of these are noted separately, but we chose to count the number of individuals who

	Kendell's tau_b		Spearman's rho			
	Correlation coefficient	Sig. (2-tailed)	Correlation Coefficient	Sig. (2-tailed)	Pearson Correlation	Sig. (2-tailed)
SUSuapp/ TEEuapp	0.599**	0.000	0.758**	0.000	0.710**	0.000
SUSuapp/ TIMEreal	-0.408**	0.000	-0.573**	0.000	-0.485**	0.000
TIMEreal/ TEEuapp	-0.490**	0.000	-0.663**	0.000	-0.595**	0.000
**Correlation is significant at the 0.01 level (2-tailed)						

Table 1. Correlations between elements of usability

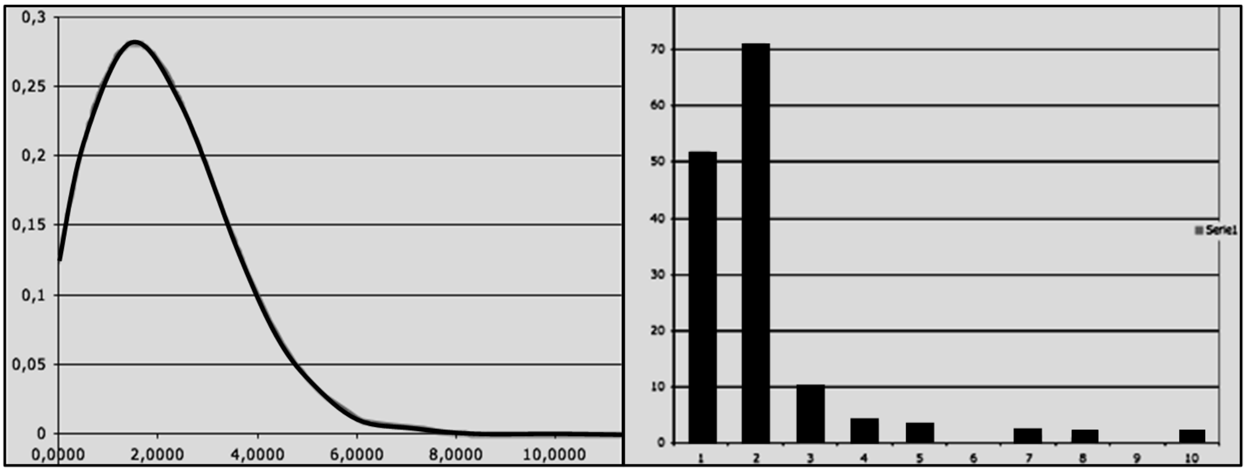

Figure 1. Rayleigh distribution and spread of times to perform use case

had experienced problems, rather than the number of problems. The seriousness of the problems could range from minor to major. However, since the ambition was to find a simple heuristic, we have not performed any qualitative analysis of the severity of the problems, but have simply noted number of users who had problems. We refer to this as No_USERS.

In this case, the cut-off point was set as being less that 33% of the total number of testers. We assume that use cases where more than 33% of users had some kind of problem are High problem, and worthy of further examination.

Table 2 illustrates the cases and their categorization as High or Low problem for Coefficient of variation and No_USERS.

We performed Fisher's exact test on the set of data shown in Table 2. This test can be used in the analysis of contingency tables with a small sample. It is a statistical test that is used to determine if there are non-random associations between two categorical variables. The results of performing Fisher's exact test are shown in Table 3.

Since the values given by Fisher's exact test are below 0.05 they can be regarded as significant, meaning that there is a statistically significant association between Coefficient of variation and No_USERS as we have defined them.

As can be seen in table 3, all of the cases (5) where No_USERS indicated a high rate of problems are discovered by the coefficient of variation being high. On the other hand, a high coefficient of variation also points to just as many cases that do not have a high rate of problems. However, the results still show that a number of use cases (8) can, with high probability, be excluded from the testing process, allowing for more efficient use of testing resources. Simply by calculating

Case No.	1	2	3	4	5	6	7	8		
Coefficient of variation	L	L	L	L	L	L	L	L		
Severity	L	L	L	L	L	L	L	L		
Case No.	9	10	11	12	13	14	15	16	17	18
Coefficient of variation	H	H	H	H	H	H	H	H	H	H
Severity	H	H	H	H	H	L	L	L	L	L

Table 2. Use cases and their categorization as High or Low problem

		No_USERS		
		High Problem	Low Problem	Total
Coefficient of variation	High Problem	5	5	10
	Low Problem	0	8	8
Total		5	13	18
		Exact Sig. (2-sided)	Exact Sig. (1-sided)	
Fisher's Exact Test		0.036	0.029	

Table 3. Coefficient of variation * No_USERS & Fisher's exact test

the coefficient of variation, 8 of 18 cases could be excluded from more expensive testing.

To conclude, the SDS records the fact that the test leader observed that users experienced problems when performing use cases, and there is found to be an association between the use cases where a larger proportion of users experienced problems, and those use cases with a high coefficient of variation. This suggests that it is possible to identify potentially problematic use cases simply by measuring the time taken to perform use cases and analysing the distribution of those times.

This article is based on research that was performed previous to the cessation of activities in UIQ. The limited number of tests that were available to be included for analysis in this study, the fact that the testing as it was performed was not designed as an experiment with this purpose in mind, and that this is a post factum analysis mean that the results must be read with some caution. However, the results we have obtained from this analysis do indicate that this is an interesting area to study more closely.

This means that it may be possible to formulate a “time it and know” formula that can be tested in new trials. This could be used to give a “problem rating” to individual use cases that could categorize the degree of problems that the user experienced. It would allow a simple categorization of use cases without needing the presence of the test leader, simply by measuring the time taken to perform the use cases, in order to identify the areas where test leader resources should be directed, thus allowing a more efficient use of testing resources.

8. Discussion

The aims of this study have been twofold: to examine the correlation between the different aspects of usability (Effectiveness, Efficiency and Satisfaction) to find whether there is one simple measurement that would express usability, and; to discover if it is possible to streamline the discovery of problematic use cases through a statistical analysis of task-completion time, which would allow scarce testing resources to be concentrated on problematic areas.

The analysis detailed above shows that, for the material collected in our study, the correlations between the factors of usability are not sufficiently strong to allow us to base usability evaluations on the basis of one single metric. This means that it is important that all three factors are measured and analysed, and as discussed previously, the test leader is an important figure in this process. This supports previous work that stresses the importance of measuring all of these aspects. This was stated to be the case even by those researchers who found stronger correlations between the different aspects measured.

However, we do find that it may be possible to discover potentially problematic use cases by analysing the distribution of use case completion times. This would mean that it is possible to collect data which indicate which use cases are most important to concentrate testing resources on. This could be done without without the presence of a test leader. Many companies involved in developing and producing mass-market products already have a large base of testers and customers who participate in different ways in evaluating features and product solutions. By distributing trial versions of software to different user groups, and by using an application in a mobile phone that measures use case completion time, and submits this data to the development company, it should be possible to collect data in a convenient manner. The development company could distribute instructions to users and testers, who could perform use cases based on these instructions, and the telephone itself could transmit data to the company, which could form the basis of the continued analysis and testing process. This data would be especially valuable since it could be based more on the use of the telephone in an actual use context, rather than in a test situation.

From an analysis of the distribution of completion times it is thus possible to gain indications of problem areas that need further attention. However, it is impossible to say, simply by looking at the completion times, what the problem may be. To discover this, and to develop design suggestions and solutions, it is still necessary for the test leader to observe and analyse the performance of the use cases that are indicated as problematic.

Future work would be to test the findings made here, by performing further tests on a greater number of devices, and comparing the results with UTUM testing as it is normally performed. It is also possible to study cases where the statistics indicate that there are problems, and other devices where this was not apparent in the statistics, and compare the results. Further work would also be to test the heuristics used in our analysis, to find if there are more accurate ways of distinguishing between low and high problem use cases.

9. Acknowledgements

This work was partly funded by The Knowledge Foundation in Sweden under a research grant for the software development project “Blekinge – Engineering Software Qualities”, www.bth.se/besq. We would like to thank Associate Professor Claes Jogréus for reading our work and giving advice regarding statistical methods, and all of our partners from UIQ Technology AB, for their help and cooperation.

References

[1] *ISO 9241-11 (1998): Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 11: Guidance on Usability*, International Organization for Standardization Std., 1998.

[2] E. Frøkjaer, M. Hertzum, and K. Hornbæk, “Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated?” in *Conference on Human Factors in Computing Systems*, vol. Proceedings of the SIGCHI conference on Human factors in computing systems. The Hague, Netherlands: ACM Press, 2000, pp. 345–352.

[3] J. Sauro and E. Kindlund, “A method to standardize usability metrics into a single score,” in *CHI 2005*, ser. Proceedings of the SIGCHI conference on Human factors in computing systems. Portland, Oregon, USA: ACM Press, 2005, pp. 401–409.

[4] J. Winter, K. Rönkkö, M. Ahlberg, M. Hinely, and M. Hellman, "Developing quality through measuring usability: The UTUM test package," in *ICSE 2007*, ser. 5th Workshop on Software Quality, at ICSE 2007, 2007.

[5] J. Winter, K. Rönkkö, M. Ahlberg, and J. Hotchkiss, "Meeting organisational needs and quality assurance through balancing agile & formal usability testing results," in *CEE-SET 2008*, ser. Preprint of the third IFIP TC2 Central and East European Conference on Software Engineering Techniques, Z. Huzar, J. Nawrocki, and J. Zendulka, Eds., Brno, 2008.

[6] J. Winter and K. Rönkkö, "Satisfying stakeholders' needs - balancing agile and formal usability test results," *e-Informatica Software Engineering Journal*, vol. 3, no. 1, p. 20, 2009.

[7] J. Winter, "Measuring usabiilty - balancing agility and formality," Licentiate Thesis, Blekinge Institute of Technology, 2009.

[8] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research," *International Journal of Human-Computer Studies*, vol. 64, no. 2, pp. 79–102, 2006.

[9] B. Pettichord, "Testers and developers think differently," *STGE magazine*, vol. Vol. 2, no. Jan/Feb 2000 (Issue 1), 2000. [Online]. Available: http://www.io.com/~wazmo/papers/testers_and_developers.pdf

[10] D. Martin, J. Rooksby, M. Rouncefield, and I. Sommerville, ""Good" organisational reasons for "Bad" software testing: An ethnographic study of testing in a small software company," in *ICSE '07*. Minneapolis, MN: IEEE, 2007.

[11] D. Schuler and A. Namioka, *Participatory Design - Principles and Practices*, 1st ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1993.

[12] J. Brooke, "SUS: A quick-and-dirty usability scale," 1986.

[13] C. Robson, *Real World Research.* Oxford, England: Blackwell Publishing, 1993, vol. 2.

[14] J. R. Lewis and J. Sauro, "The factor structure of the system usability scale," in *LNCS 5619*, vol. Proceedings of the human computer interaction international conference (HCII 2009),. Springer Verlag, 2009, pp. 94–103.

[15] T. S. Tullis and J. N. Stetson, "A comparison of questionnaires for assessing website usability," 2004. [Online]. Available: http://home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf

[16] BTH, "UIQ, usability test," Aug. 2008. [Online]. Available: http://www.youtube.com/watch?v=5IjIRlVwgeo

[17] Y. Dittrich, K. Rönkkö, J. Erickson, C. Hansson, and O. Lindeberg, "Co-operative method development: Combining qualitative empirical research with method, technique and process improvement," *Journal of Empirical Software Engineering*, vol. 13, no. 3, pp. 231–260, 2007.

[18] R. K. Yin and S. Robinson, *Case Study Research – Design and Methods*, ser. Applied Social Research Methods Series. Thousand Oaks, Cal.: SAGE publications, 2003, vol. 3.

[19] B. G. Glaser and A. L. Strauss, *The discovery of grounded theory : strategies for qualitative research.* Piscataway, NJ.: Aldine Transaction, 1967.

[20] K. Rönkkö, "Ethnography," in *Encyclopedia of Software Engineering (accepted for publication)*, P. Laplante, Ed. New York: Taylor and Francis Group, 2010.

[21] J. Dumas and J. Redish, *A Practical Guide to Usability Testing.* Exeter, England: Intellect, 1999.

[22] G. Denman, "The structured data summary (SDS)," 2008.